REPORT DOCUMENTATION PAGE	Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY) 26-08-2015	2. REPORT TYPE Final Report	3. DATES COVERED (From - To) 1-Jun-2012 - 31-May-2015

4. TITLE AND SUBTITLE Final Report: Electroencephalogy (EEG) Feedback In Decision-Making	5a. CONTRACT NUMBER W911NF-12-1-0213
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER 611102
6. AUTHORS Matthew S. Peterson	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES George Mason University 4400 University Drive, MSN 4C6 Fairfax, VA 22030 -4422	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES) U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	10. SPONSOR/MONITOR'S ACRONYM(S) ARO
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) 61763-NS.1

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

The goal of this project is to investigate whether Electroencephalogy (EEG) can provide useful feedback when training rapid decision-making. More specifically, EEG will allow us to provide online feedback about the neural decision processes occurring during image-recognition training, and in turn will lead to faster decision responses, more hits, and fewer false alarms. A second subgoal is to investigate whether this can be done under free-viewing conditions in which eye movements must be made while viewing complex scenes.

15. SUBJECT TERMS

Brain-Computer Interface, Training, Decision Making

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Matthew Peterson
a. REPORT UU	b. ABSTRACT UU	c. THIS PAGE UU	UU		19b. TELEPHONE NUMBER 703-993-4255

Report Title

Final Report: Electroencephalogy (EEG) Feedback In Decision-Making

ABSTRACT

The goal of this project is to investigate whether Electroencephalogy (EEG) can provide useful feedback when training rapid decision-making. More specifically, EEG will allow us to provide online feedback about the neural decision processes occurring during image-recognition training, and in turn will lead to faster decision responses, more hits, and fewer false alarms. A second subgoal is to investigate whether this can be done under free-viewing conditions in which eye movements must be made while viewing complex scenes.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received Paper

TOTAL:

Number of Manuscripts:

Books

Received Book

TOTAL:

Received Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Eric J. Blumberg	0.50	
Mellisa Smith	0.50	
FTE Equivalent:	**1.00**	
Total Number:	**2**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Matthew S. Peterson	0.12	
FTE Equivalent:	**0.12**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:...... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):...... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:...... 0.00

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

Technology Transfer

Final Report

"Electroencephalogy (EEG) Feedback In Decision-Making"

August 26, 2015

Issued by

Army Research Office
IMOPAT ATO

Contract No. W911NF-12-1-0213

Dr. Matthew S. Peterson, Principal Investigator
George Mason University
Psychology Department
ARCH lab
Fairfax, VA 22030-3f5
703-993-4255

Reporting Period 6/1/11 - 5/30/15

Abstract

The goal of this project is to investigate whether Electroencephalogy (EEG) can provide useful feedback when training rapid decision-making. More specifically, EEG will allow us to provide online feedback about the neural decision processes occurring during image-recognition training, and in turn will lead to faster decision responses, more hits, and fewer false alarms. A second subgoal is to investigate whether this can be done under free-viewing conditions in which eye movements must be made while viewing complex scenes.

Statement of purpose

Our approach to exploring whether an Electroencephalogy-driven Decision Aid (EEG-DA) can be used to enhance training has two broad steps. The first step is to develop a classifier and to determine whether is can enhance simple perceptual recognition (presentation of a single stimulus at a time) and whether this ability transfers to other stimuli. The second step is to determine whether this can be scaled-up to more complex tasks and environments, such as visual search.

Timeline

Year 1. **Develop Classifier:** Our first goal is to develop an EEG-Based classifier for identifying whether an object is recognized or not under laboratory conditions. This task involves several parts, including (1) finding a suitable machine-learning algorithm (e.g. support-vector machine, linear discriminate analysis, etc.), (2) determining the minimal number of electrodes to minimize set-up time while maximizing classification, (3) finding the optimum experimental design to balance task realism with clear signals for classifying EEG signals.

Year 2. **Evaluate Participant Training**. Our main concern is whether an EEG-DA can be used to enhance perceptual training. One of our original goals was to determine if feedback could *decrease speed* and *increase accuracy* of classification responses. However, based on pilot results from Year 1, measuring speeded responses will not be possible due to necessary changes in the training paradigm. Instead of focusing on response speed, we now will focus on enhancing accuracy as stimulus display time (and information) is reduced.

Year 3 (past year). **Continue to Evaluate Participant Training.** During Year 2, we found huge variability in individual differences in regard to EEG classification, and these differences are unrelated to behavioral results (i.e. how well subjects performed the task). In year 3 we attempted to develop a behavioral prescreener that might predict which participants would be easy to classify, but that was unsuccessful. Instead, we ran participants in the training study, and determined whether they received sham or EEG-DA feedback the second session based on how successful we were at classifying their EEG waveforms from the first session.

Throughout this document we will be using the term *training* to refer to the process by which participants become better at our perceptual identification task though experience with the task and stimuli. So as to prevent confusion, when possible we will use the word *tuning* to refer to training the pattern-recognition classifier to recognize when an individual is viewing a target, lure, or empty scene based on the participant's EEG/ERP signals.

Year 1

Task Evaluation (RSVP vs. single shot)

Our original approach was to mimic the stimulus presentation technique used by *ROC-V*, which we will call the one-shot approach. We chose the one-shot approach because it is not only similar to the current *ROC-V* training used by war-fighters, but also because it mimics what an observer would see in a single glance. In addition, some of the original ERP research investigating object classification (Johnson & Olshausen, 2003) used this approach. Later work (Luo & Sajda, 2009) comparing the one-shot approach (*CSVP-Continuous Serial Visual Presentation*, in the authors' terminology) to the RSVP (Rapid Serial Visual Presentation) approach traditionally used in single-trial BCI (Brain-Computer Interface) tasks suggested a similar effect-size and scalp distribution. However, we were unable to get adequate single-trial classification using the one-shot technique. Starting with Experiment 3, we switched to using the RSVP technique. Because of this switch precludes us from gathering response times, we will be switched our training approach to emphasize increased accuracy under situations of limited information

Equipment

The EEG was recorded using a Neuroscan NuAmps amplifier. Recordings were made at 32 scalp sites (extended 10-20 system) with Ag/AgCl electrodes mounted in an elastic cap. In addition, Ag/AgCl electrodes were placed at the left supraorbital and suborbital sites, as well as the left and right outer canthal sites, to monitor vertical and horizontal electro-oculographic (EOG) activity, respectively. Customized MATLAB scripts (The MathWorks, Natick, MA) and the Psychophysics Toolbox were used to present the stimuli, while SCAN 4.01 software (Compumedics, North Carolina, USA) was used to digitize the EEG at a sampling frequency of 500 Hz. All scalp electrodes were referenced to the left mastoid on-line and re-referenced to the average of the left and right mastoids following data collection. All electrode impedances were maintained below 5 kΩ and recorded with a 70 Hz low-pass filter. All machine learning was done offline using BCILab.

Experiment 1 Our first experiment is based on the cued-target methodology of Johnson & Olshausen (2003). The target stimuli were six different armored vehicles (T-72, 2S3, BMP-1, BTR-60PB, BTR-70, Mk 13 Centurion) taken from the ROC-V image database (courtesy of RDECOM:NVSED). Lures consisted of the following six armored vehicles from the ROC-V image database: 2S1, LAV LOG, LAV-25, M113, Pz68, and T-34. All images were taken from ground level (ROC-V also includes aerial views), and eight different viewing aspects (45°) were used for each vehicle. All images were converted to grayscale.

The task for the participant was two-alternative-forced-choice (2afc), with cued recognition. Prior to the start of each trial, the participant was presented with the name of a cued target (e.g. "Is this a BMP-?"). This remained on the screen for 5 seconds or until the subject pressed the spacebar, whichever came first. Next, the participant was presented with either an image of a lure or an image of the cued target. If the cued target was present, the subject pressed the "/" key, and when cued target was absent, the subject pressed the "z" key. Subjects were instructed to respond as quickly and as accurately as possible. Each experimental session consisted of 192 trials, half of which required 'present' responses.

All three participants had trained with the ROC-V mobile app to 95% accuracy before participating.

Results

Behavioral results averaged 90% correct, indicating that participants could clearly differentiate the targets from the lures.

For classification of ERP data, we tried two machine-learning algorithms: support-vector machine (SVM) and Linear Discriminant Analysis (LDA). For analysis, we used the windowed-means paradigm, in which the input vector for a given trial consists of the mean amplitude (for each electrode) within a given time window(s). The best classification occurred with LDA, using 6 windows (from 100 to 600 msec post-stimulus) and all 32 scalp electrodes. Table 1 shows the best and worst classification using 5-fold cross-validation:

Table 1

	Hit %	FA %	A'	d'
Best	76%	54%	.61	0.61
Worst	64%	49%	.55	0.25

Hit % is the percentage of correct classifications when the target was present.

FA% (False Alarm) is the percentage of incorrect classifications when a target is absent.

A' is a non-parametric estimate of the area under the ROC curve, and is a measure of sensitivity that varies between 0 and 1, with 1 being perfect performance

d' is a Signal Detection Theory measure of sensitivity measured in z-space.

Although it was encouraging that our hit rates were above chance (50%), our observed false alarm rates hovered around chance. The low A' and d' scores reflect poor single-trial classification. Visual inspection of the average waveforms suggested clear differences between the target and lure trials in the N200 and P300 region, but the differences were not large enough for adequate single-trial classification.

Experiment 2

One possible confound in Experiment 1 is that by requiring subjects to respond as quickly as possible, motor-driven waveforms might have added noise to the EEG data. To prevent motor components from potentially contaminating the data, we used a prompted task, in which subjects were first shown an image for 5 seconds followed by a prompt (e.g. "Is this a BMP-1?"). Subjects were instructed to respond as quickly as possible to the prompt by pressing the "z" key for correct and the "/" key for incorrect. Because we were interested in recognition of learned targets, all analyses were time-locked to image presentation (not to the response).

Results

Although our behavioral results remained excellent (>90% correct), single-trial waveform classification remained poor, as can be seen in the Table 2. As in Experiment 1, LDA provided the best classification among our subjects.

Table 2

	Hit %	FA %	A'	d'
Best	64%	54%	.55	0.26
Worst	86%	80%	.53	0.24

Experiment 3

Our first two pilot experiments yielded very poor single-trial classification results. We originally chose the single-shot methodology as it most closely mimicked a conventional training situation. Although nearly all BCI paradigms have used a variant of the RSVP technique, there was no indication in the literature as to why this was chosen. Indeed, several papers had suggested that the choice was because it maximized image throughput, and a few traditional studies suggested that a single-shot approach would be as effective as RSVP (Johnson & Olshausen, 2003; Luo & Sajda, 2009).

An investigation of the neuroscience literature outside of BCI suggested that single-trial classification might be leveraging the attentional-blink (AB) driven p300 effect when the items are presented in RSVP. In RSVP, items are flashed one at a time to the same spatial location and at a rapid rate (typically 100msec per item presentation rate). The attentional blink (Chun & Potter, 1995; Raymond, Shapiro, & Arnell, 1992) occurs when an individual has to report two targets that appear somewhere in an RSVP stream. The second target is more likely to be missed if it occurs in a window roughly 200-500 msec after the first target. The deficit is thought to occur because when the second target occurs, mental resources are still processing the first target. Because of the nature of the RSVP stream, there is no visual persistence, and the second target is erased before mental resources can be devoted to identifying it.

Important to us is the finding by McArthur and colleagues (1999) that the size of the attentional blink is correlated with the size of the p300 ERP waveform elicited by the first target, and that both the behavioral AB deficit (to the second target) and p300 (from the first target) become more pronounced as the first target becomes more difficult to identify. Although none of the BCI experiments we researched were designed to investigate the AB, it appeared that the AB effect might be magnifying the p300, and in effect increasing the signal-to-noise ratio.

In the following three experiments, starting with Experiment 3, we adopted a modified RSVP technique. As in Experiment 2 (and like traditional AB experiments), we had subjects respond "offline" after the RSVP stream had ended.

Common Methods for Experiments 3-5

Each trial consisted of 20 images flashed one-at-a-time for 200 msec each. Distractors consisted of 33 possible background images. Some of these background images are from the ROC-V training set (vehicles were removed from the images by NVESD), whereas the remainder came from the USDA NRCS photo gallery (http://photogallery.nrcs.usda.gov). Each target-absent trial consisted of 20 distractor frames, whereas target-present and lure-present trials (Experiments 4 and 5) consisted of 19 distractor frames and 1 target or lure frame occurring in position 5-15. An example of a target is shown in the left frame of Figure 1. The right frame of Figure 1 shows an example of a distractor image.

Experiments 3-5 consisted of 432 trials broken into blocks of 72 trials, with each block having a single cued target (e.g. "T-72"). The cue occurred at the beginning of each block of trials and consisted of the name of the vehicle as well as cropped examples of the vehicle from each aspect ratio.

For Experiment 3, one-third (144, or 24 per block) of the trials contained a target vehicle and required a target-present response. The remaining trials were all target-absent.

All reported results are 5-fold cross validation using LDA for classification.

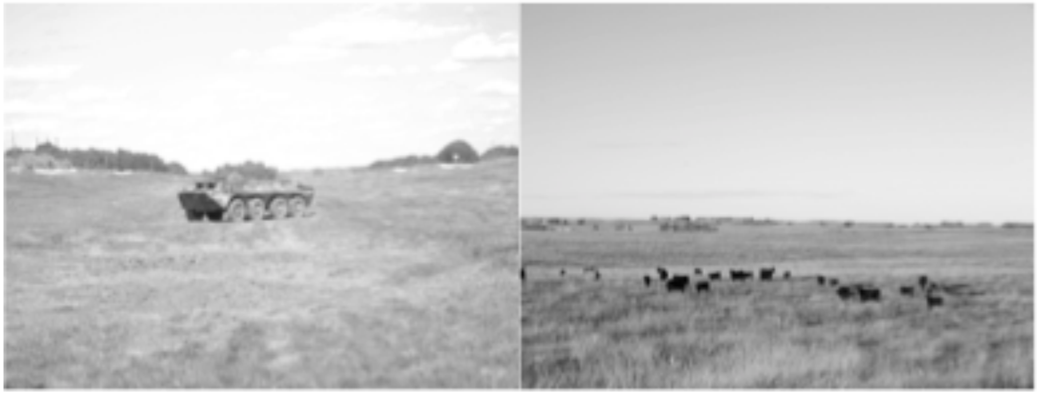

Figure 1

Results

As before, behavioral results were high and averaged over 95% for our three subjects. As can been seen in the Table 3 below, our classification rates for our 3 subjects are greatly improved. In particular, the decreased false alarm rates are encouraging, as high false alarm rates can lead users to distrust automated decision aids. (Parasuraman & Wickens, 2008).

Table 3

	Hit %	FA %	A'	d'
Best	89%	28%	.81	1.8
Worst	86%	38%	.74	1.6

Experiment 4

Although the results of Experiment 3 are encouraging, to perform the task, subjects were only required to detect the presence of a vehicle. Since our ultimate goal is to provide feedback on vehicle *identification*, this is not sufficient. Experiment 4 was changed to an identification task by introducing lures. One third of the 432 trials contained a target, one third contained a lure, and one third contained no vehicles in any of the 20 RSVP frames. Within each block of 72 trials, one third of the trials contained the cued target, one third contained a lure, and the remaining trials consisted solely of background distractor scenes. Subjects were told to respond with a keypress to indicate if a target or a lure was present, and to withhold a response if the RSVP stream consisted solely of background images.

Results

Shown in Table 4 are the best and worst 5-fold cross-validation results using an LDA classifier for discriminating between a target and a distractor. As can be seen below, our classification results are improved over Experiment 3, with both higher hit and lower false alarm rates.

Table 4

	Hit %	FA %	A'	d'
Best	92%	26%	.83	2.1
Worst	89%	29%	.80	1.8

Shown below in Table 5 are the classification rates for the targets *vs.* lures. Clearly, the classifier had a more difficult time discriminating between the ERP waveforms elicited by the target and lure trials, but so did our subjects, whose accuracy ranged from 75-85% (33% is chance). Classifier training included all target and all lure trials, without regard for accuracy, and the decreased response accuracy meant that the classifier was trained with examples that included incorrect responses.

Table 5

	Hit %	FA %	A'	d'
Best	56%	40%	.58	0.4
Worst	54%	45%	.55	0.23

Experiment 5

The lures used in Experiment 4 were chosen because they are historically known to be confusable with the targets, and we wanted to maximize task difficulty. For Experiment 5, we chose lures that were easier to discriminate from the target set, with the reasoning being that the initial training procedure would include more easily discriminable stimuli, with discrimination difficulty increasing as the trainee became more proficient. These easier lures include the 1S12 Longtrack, Bluebird Bus, Ford 6610 Tractor, YAZ-463, Z-Turn mower, and ZIL-157.

Experiment 5 was otherwise identical to Experiment 4.

Results

Shown in Table 6 are the best and worst 5-fold cross-validation results using an LDA classifier for discriminating between a target and a distractor. As can be seen below, we are able to easily discriminate between target and distractor stimuli based on ERP recordings.

Table 6

	Hit %	FA %	A'	d'
Best	96%	14%	.91	2.8
Worst	89%	22%	.84	2.0

Shown below are graphs of electrode FC3 for a single participant. The top graph (Figure 2) is for distractors and the bottom graph (Figure 3) is for waveforms in response to targets. The top part of each graph shows color intensity plots for each stimulus presentation. Along the bottom is the average waveform corresponding to the top intensity plot. The periodic waveforms due to stimulus presentation (200msec, or 5 Hz) are easily seen in the plot for the distractors.

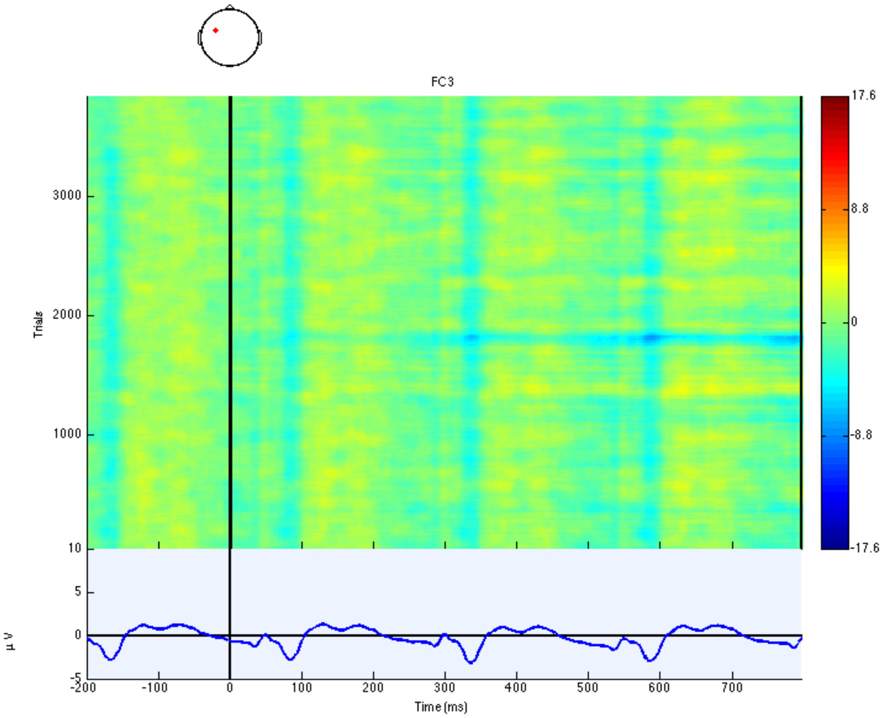

Figure 2

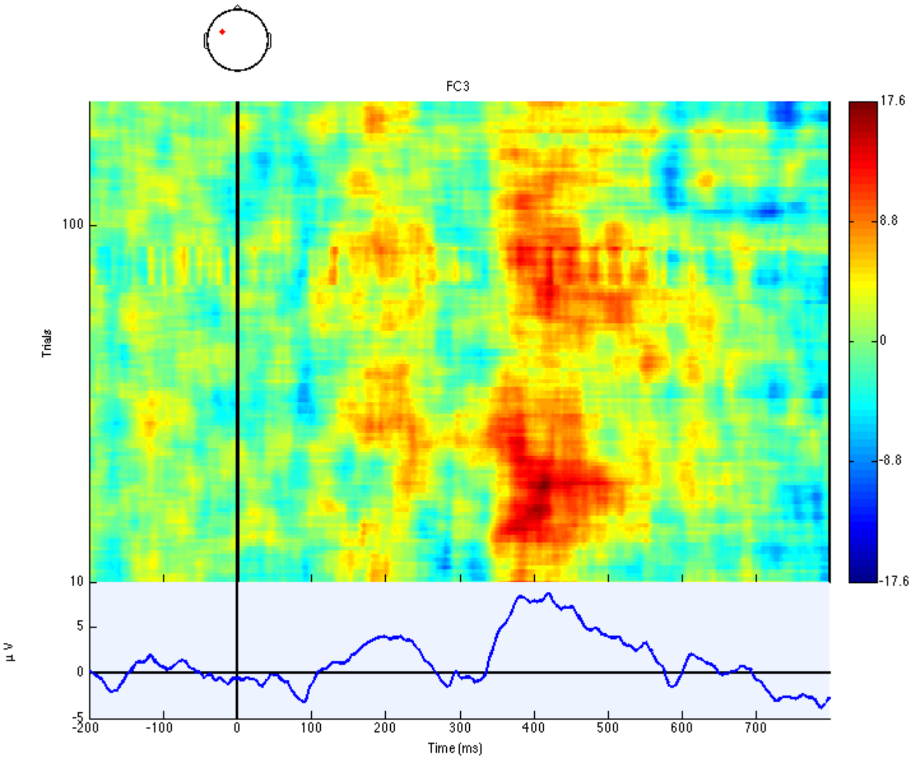

Figure 3

In comparing Figure 2 (distractors) to Figure 3 (targets), one can easily visualize the difference in waveforms, particularly in the regions peaking 200 and 400 msec post-stimulus. Shown below in Figure 4 are scalp maps of the weights for the LDA classifier for one participant.

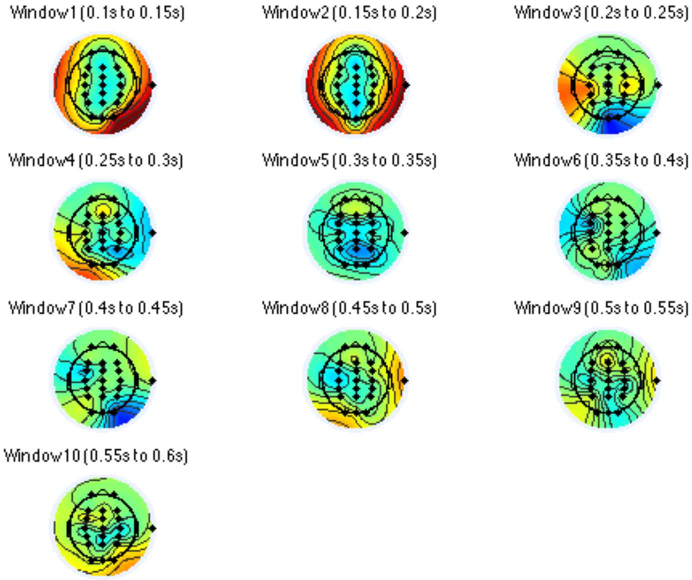

Figure 4

More important than discriminating between targets and empty backgrounds is the ability to discriminate between targets and lures. Behavioral accuracy for this task was near 100% for all subjects. Although we are satisfied with the classification rates (see Table 7), we would prefer to have a lower false alarm rate. Methods of dealing with the false alarm rate will be discussed in the Plans for Coming Year section.

Table 7

	Hit %	FA %	A'	d'
Best	79%	24%	.78	1.5
Worst	76%	31%	.73	1.2

Figure 5 shows electrode FC3 for one subject when viewing lures, and Figure 6 shows the scalp distribution of LDA weights for targets vs. lures. The largest difference between targets and lures at electrode site FC3 appear to be between 350-450 msec for this subject.

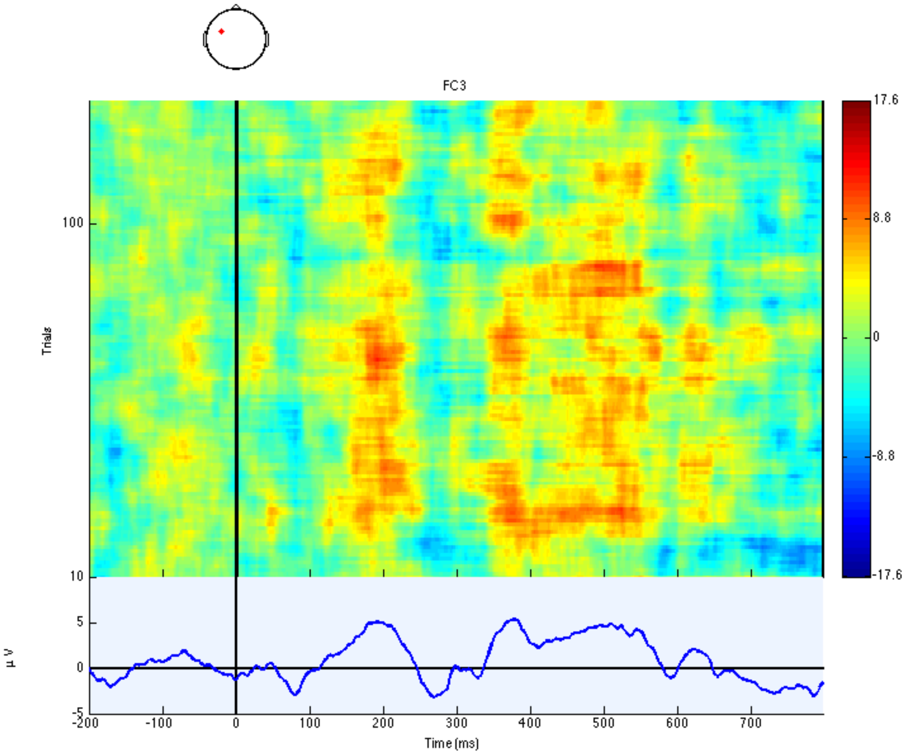

Figure 4

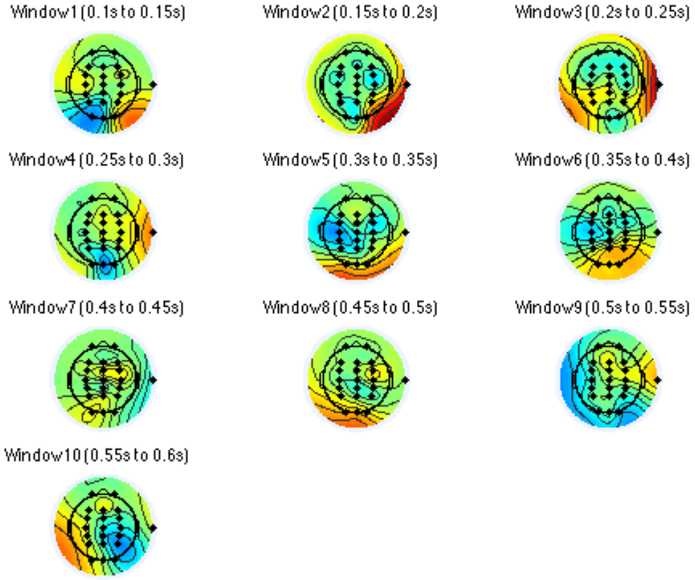

Figure 5

Electrode evaluation

Because our subjects will need to come back for multiple sessions of training, we decided to evaluate the feasibility of using fewer electrodes in order to speed up preparation time. Shown below are the maximum and minimum classification sensitivity scores across our subjects for an earlier version of our experiments. We decided that the trade-off for using fewer channels was unacceptable.

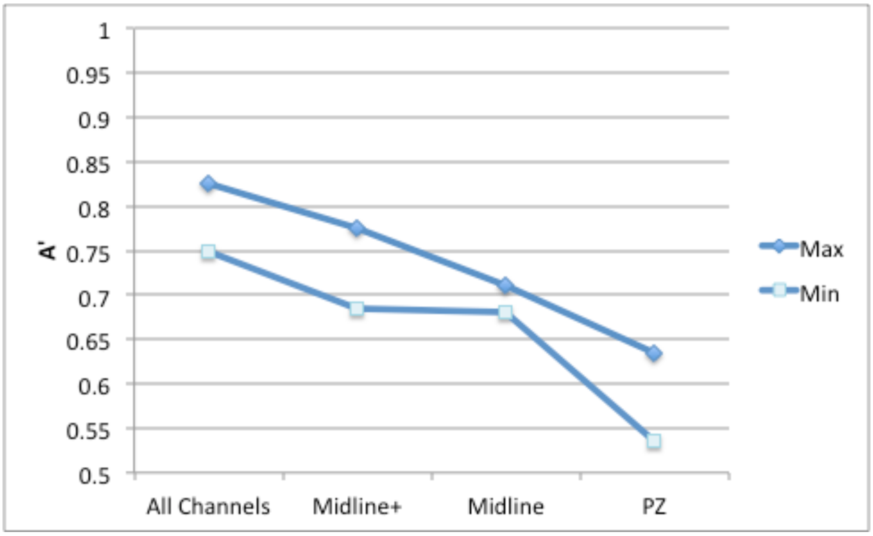

Figure 6

Tuning Evaluation

In addition to evaluating how many channels were necessary to get sufficient classification, we also evaluated how many trials would be needed to adequately tune the classifier for the training regimen. To prevent confusion, we use the term ‘tuning’ to refer to classifier optimization (this is traditionally called “training” in the literature), and “training” to refer to the learning experiments that the subjects will undergo. In other words, we train the people and tune the classifier. To do this, we first mined our datasets and later confirmed the results with naive subjects. Specifically, we first tuned the classifier on various amounts of data (the first 10%, 25%, or 50%) and tested on the remaining 50% of the data from a single session. We found that we would need roughly 216 trials (3 blocks of 72 trials, or roughly 20 minutes of data) to tune a classifier that could adequately generalize to the rest of the dataset for targets vs. lures (A’ ranging from 77-73%). The majority of misclassifications were false alarms.

Programming

One of the hurdles we had to overcome was how to provide feedback in real-time to the participants. With help from employees at MITRE, we adapted their code for the BrainVision system so that it will work with our Neuroscan system. This code allows the Neuroscan Scan software (running on a host PC) to send EEG data and markers over TCP/IP to a MacBook Pro hosting Matlab and BCILab. In the case of EEG feedback condition, the MacBook classifies the EEG data in realtime and sends the classification results to the host computer at the end of each trial. For the sham feedback condition, the Macbook pseudo-randomly generates feedback when it receives an end-of-trial marker.

Year 2

Software Development

Although we are using BCI-Lab for developing our classifier, one problem was that with our system (Neuroscan NuAmps), BCI-Lab only allowed offline processing and did not provide a plug-in for the real-time streaming and analysis of data from our EEG system. We developed a series of Matlab Neuroscan plug-ins for BCILab and submitted the open source code to the BCILab depository.

Training Experiments

EEG Training Experiment 1

Before starting the feedback-training experiment, we wanted to establish several baselines. These include:

• *Learning Curves*: Without feedback, how quickly do subjects learn the training stimuli? This is important because we need a task that is sufficiently difficult so that feedback-training has room to potentially improve performance.

• *Day-to-day classification*: How much day-to-day variability is there in classifying the EEG signals? This is important, as too much variability (e.g. high-quality feedback one day, poor quality the next) could affect the trainee’s trust in the automated feedback (Parasuraman & Wickens, 2008).

• *Classification Generalizability*: How well would a classification solution derived from an early session work for later sessions? Because classification needs a large body of data, it would be ideal if a classification solution from the first day of training could be used on subsequent days, with at most only minor updating each day. The alternative would mean that before feedback-training could occur, each session would require extra blocks of trials to gather data to tune the classifier.

We ran 4 naive subjects over 4 sessions, while gathering both behavioral and EEG data during each session. Each session contained 432 trials, with one third of the trials containing a target, one third containing a lure, and one third containing no vehicles in any of the 20 RSVP frames, with each frame presented for a duration of 200 msec. Within each block of 72 trials, one third of the trials contained the cued target, one third contained a lure (non-target vehicle), and the remaining trials consisted solely of background distractor scenes. Each block of trials had the same target vehicle (e.g. T-72), and the vehicle could be presented in any of eight aspects (rotations). Figure 1 shows examples of target and background scenes. Subjects were told to respond with separate key presses to indicate if a target or a lure was present, and to withhold a response if the RSVP stream consisted solely of background images. To become familiar with the targets, on the first day subjects were allowed to study the targets while the EEG cap was being fitted (approximately 20-30 minutes) using *ROC-V Mobile* on an Android phone.

Figure 7 shows the behavior results over the 4 sessions measured by A', an unbiased measure of accuracy that takes both hits and correct rejections into account. Overall accuracy is quite high and near ceiling, indicating that future experiments will require either a faster presentation rate or a shorter presentation duration in order to degrade performance and make the task more difficult. This near-ceiling performance by the naïve subjects surprised us, as all prior pilot subjects had extensive experience with the stimuli, including one subject who had written and debugged the presentation software as well as another subject who was involved in editing the images.

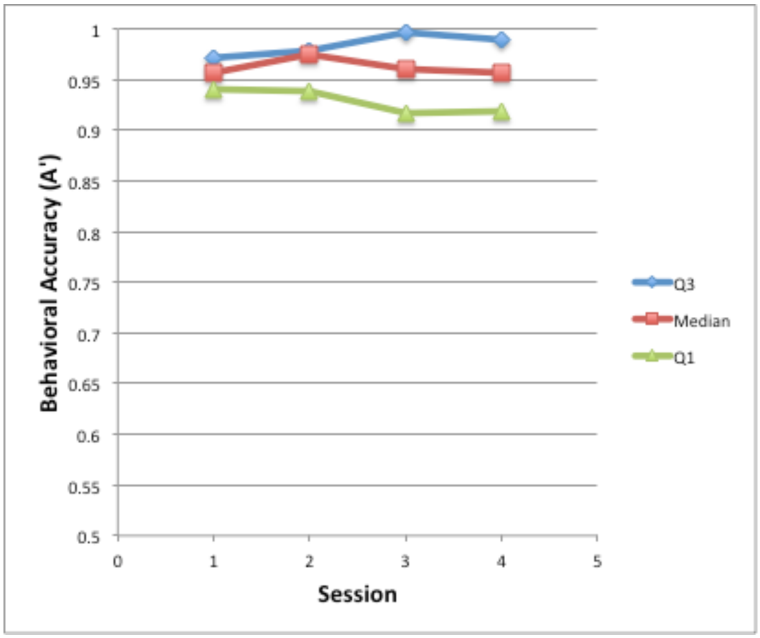

Figure 7

Figure 8 shows the EEG classifications averaged over the four subjects for each session. Specifically, these are the grand average 5-fold cross-validation results using an LDA classifier for discriminating between targets and a distractors or a targets and lures. Overall, our classification accuracies are much lower than expected. Specifically, Tables 6 and 7 from

Experiment 5 from Year 1 show classification (A') rates that ranged between 0.91 and 0.84 for targets vs. distractor backgrounds and between 0.78 and 0.73 for targets vs. lures.

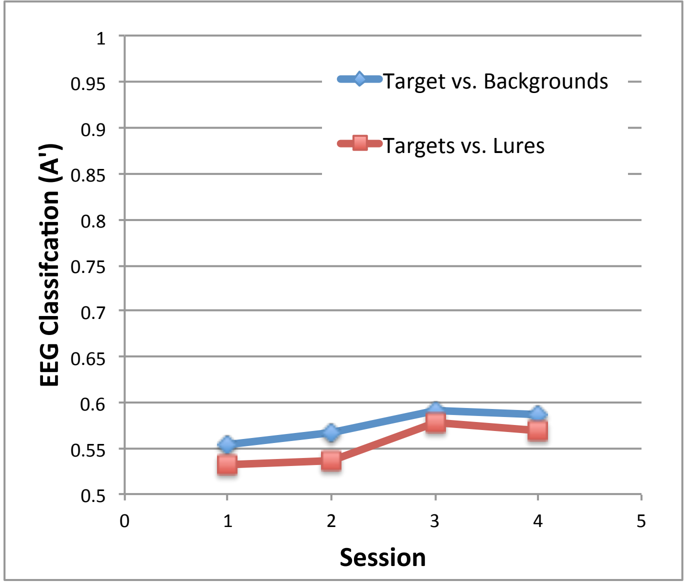

Figure 8

Training Experiment 1 (Year 2) and *Experiment 5* (Year 1) were identical in methodology, with the only difference being that the 3 subjects participating in Experiment 5 had prior extensive experience with the stimuli. Clearly behavioral accuracy is not a good predictor of classification accuracy, as the subjects in the current experiment performed near ceiling in the first session. We hypothesized that the superior classification rates might be due to over-training with the stimuli, and tested that in *Training Experiment 2* by over-training two naïve participants.

In addition, we also examined how well the classification solutions generalized from day-to-day. Figure 9 shows the average classifications for all four days, with each grouped line representing a separate day of training. The first point on each line (day one has only one point) illustrates the average classification based on that session's EEG data, and is identical to the target vs. backgrounds data in Figure 8. Each subsequent point shows how a classifier trained using a previous session's data classified the current session. For example, the second point from the left (blue circle) shows how well a classifier trained on session two's data classified the data from the same session. The third point (orange square) shows how well a classifier trained on day 1 classified data from day 2. Two overall effects are apparent. First, classification accuracy is overwhelming determined by the session, with later sessions (i.e. days 3 and 4) showing much better classification than the earlier sessions (this can also be observed in Figure 8). The second overall effect is that the data set that the classifier was trained on does not appear to matter. For example, the four points on the right-hand side of the graph all represent classification accuracy for the data from session 4. In comparing those four points, it can be seen that classifiers tuned with data from days 1-3 worked as well as a classifier trained with data from day 4. Although overall classification is poor, it nonetheless suggests that we might be possible to use classifiers from previous sessions for the feedback experiment without having to retrain (tune) the classifier for each session.

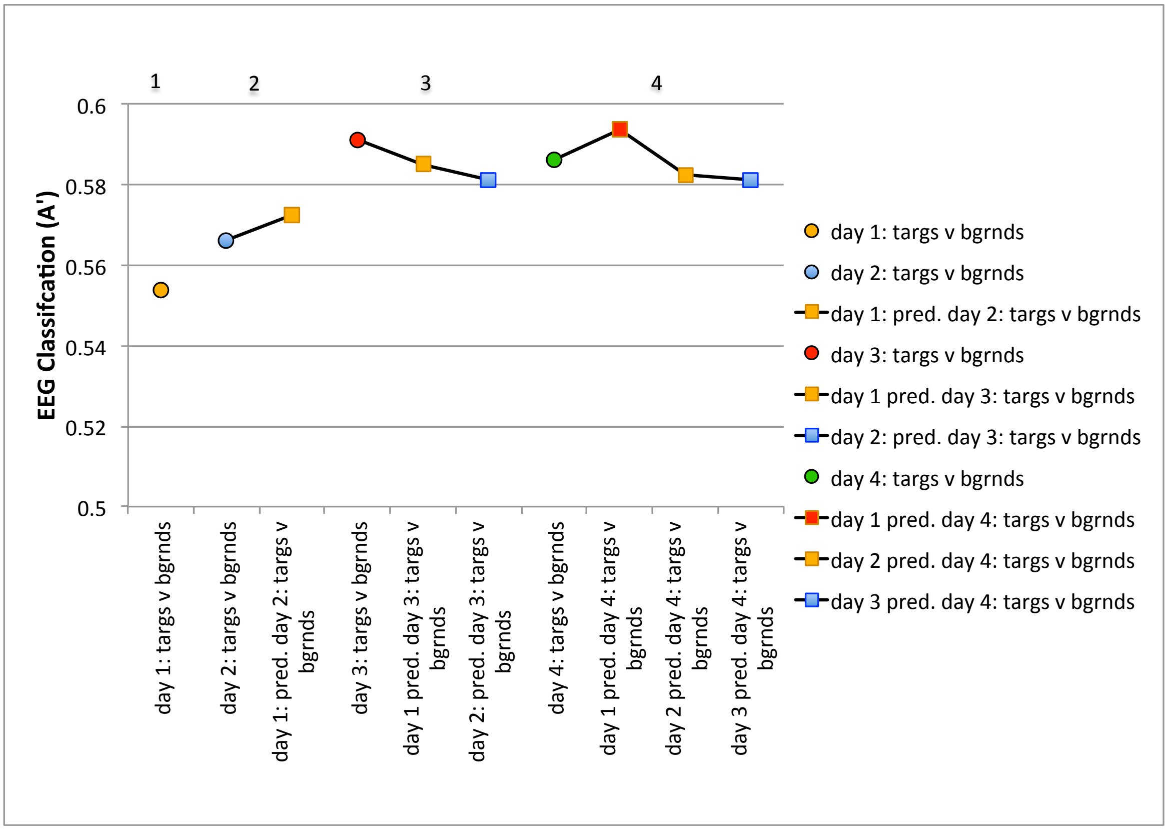

Figure 9

EEG Training Experiment 2

To determine whether over-training was the cause of our previously higher classification rates, two naïve subjects trained for four days on the stimulus set. When tested in the RSVP experiment on the fifth day, both achieved behavioral performance close to or at ceiling: A' of 0.98 and 1.00. However, their EEG classification was markedly lower: A' of 0.55 and 0.61 for discriminating between target and distractor backgrounds, and A' of 0.54 and 0.53 for targets vs. lures. This suggests that overtraining is not a factor in classifier performance, but instead classifier performance might be due to individual differences.

Attentional Blink Pre-screener

Individual differences in brain morphology, skin conductance, and skull shape can all lead to varying levels of signal-to-noise in regard to ERPs. The goal of the next experiment was to see if we could devise a quick behavioral pre-screener that would allow us to cherry-pick subjects that are likely to show good EEG/ERP classification.

One candidate for a behavioral marker is the *attentional blink* (AB, Chun & Potter, 1995). The attentional blink can occur when trying to identify two targets presented in an RSVP stream. Specifically, the second target is often missed when it occurs in a window 200-500 msec after the first target. This occurs because the second target is presented while the first target is being mentally processed. Because the stimuli are presented using RSVP, this prevents visual persistence and the second target is masked by later stimuli.

For our purposes, the size of the attentional blink appears to be correlated with the size of the p300 ERP component (McArthur et al., 1999), which, based on last year's data, appears to be one of the components that classifiers are sensitive to in our task. However, what want to know is slightly different: does the size of the attentional blink correlate with the strength of EEG classification? In the attentional blink p300 literature, the EEG signals are time-locked to the presentation of either the first or second target, which is identical to what are doing when we time-lock our classification signals to the stimulus presentation in our RSVP training experiments.

To this end, we replicated the experiment of Sessa et al. (2007). In their experiment they presented a series of RSVP streams containing white letters on a gray background at the rate of 83 msec per item. The first target (T1), when present, was a black numeral. The second target (T2) was the letter 'E'. Participants' goal was to identify the black numeral and report whether an 'E' had appeared or not. A key to this experiment was that when the 'E' was present (20% of the time), it was either at lag 3 (two intervening items, or 250 msec after the first target) or at lag 7 (581 after the target). When T2 is presented at lag 3, it is in the prime spot for the AB to occur. In their experiment, T2 was either the last item in the stream, meaning it was not masked by a following item (unmasked), or it was the penultimate item in the RSVP stream (masked). In their experiment, the AB, as measured by the T2 accuracy difference between lag 7 and lag 3, was larger for the more difficult masked T2.

We ran five participants in the AB experiment, and achieved a nice range of behavioral results, with the magnitude of the attentional blink varying between 0 and 0.7 for the unmasked condition (mean = 0.29) and between 0 and 0.9 for the masked condition (mean = 0.23). Unfortunately, we found no correlation between the size of the AB and classification accuracy (see Figure 10). We discuss our plans for future subject screening in the next section.

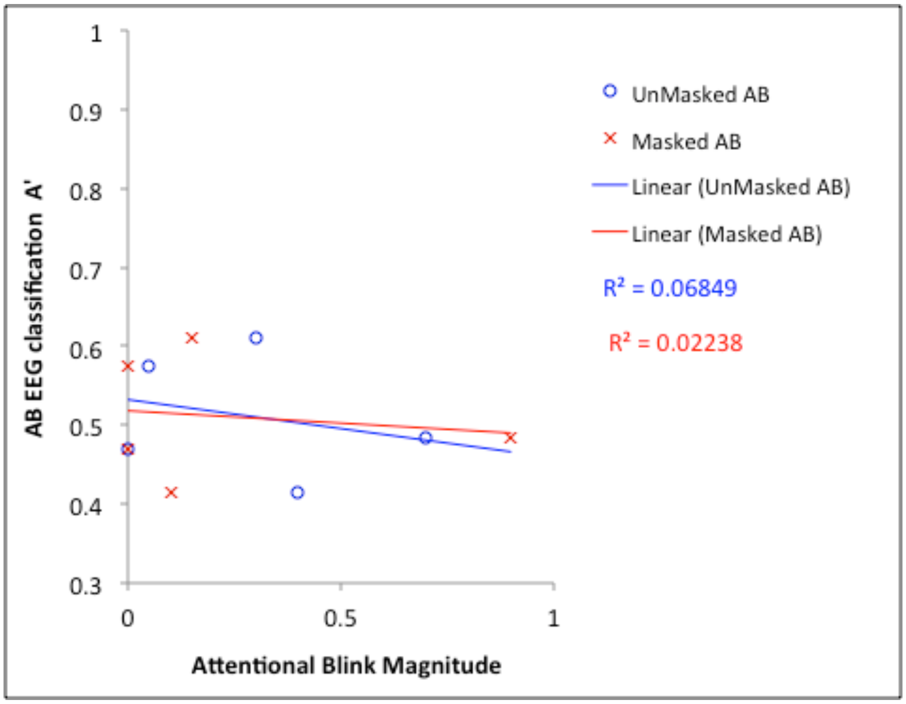

Figure 10

Year 3

EEG Training Experiment 3

Method

EEG training Experiment 3 was the first of our studies to examine how giving feedback to naïve participants would potentially enhance training over multiple sessions. Unlike our previous experiments, which used lab members as pilot subjects, Training Experiments 3 and 4 used participants from the George Mason University undergraduate Psychology subject pool. Both experiments were 3 days sessions, with feedback (sham or EEG-DA) occurring on the second and third days.

EEG recordings were made on all three days, and like previous experiments, participants were allowed to study the test vehicles using ROC-V Mobile on an Android phone during EEG cap setup on the first day. Given the large variability in our ability to classify individuals' EEG waveforms, we decided to triage subjects based on how well we could classify their EEG based on the first day's data. Given our previous low classification rates, we decided to initially set a low threshold (A' = 0.7) for inclusion in the EEG-DA feedback group (see Conclusions section).

Because our subjects were undergraduate students naïve to the stimuli and task, we decided train them using the 200 msec display time used in *Training Experiment 1*. However, as can be seen in Figure 11, this task was too easy and subjects' performance was near ceiling. Because of this, we decided to switch to a faster presentation rate (100 msec/item) for *Training Experiment 4*. Because none of the participants' EEGs were classified with an A' > .7 on the first day, the participants in this experiment all received sham feedback on days 2 and 3.

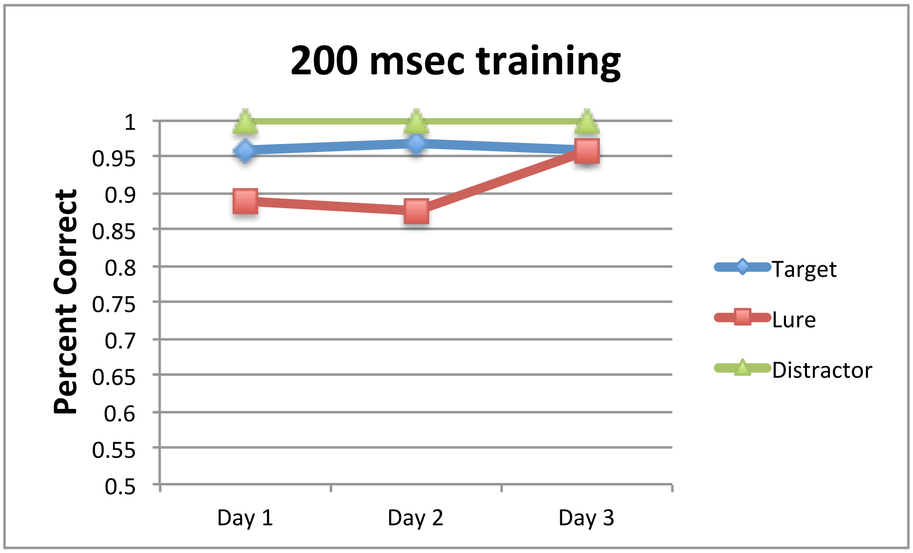

Figure 11

EEG Training Experiment 4

EEG Training Experiment 4 was identical to the previous experiment, except that the presentation duration of each image was shortened from 200 to 100 msec. In addition, we decided to swap the lures for a more difficult set of armored lures (e.g. T-34 tank instead of Bluebird Bus), as we thought that Experiment 3 might have become a presence/absence judgment (armored vehicle present or not). As can be seen in Figure 12, this had the benefit of bringing the behavioral performance down to an acceptable level (70-75% accuracy) that would allow potential training benefits to be observed. Unfortunately, we had two problems occur that prevented us from collecting an adequate amount of data before the end of the final reporting period. The first issue was an intermittent bug in the feedback software that would cause a crash midway through an EEG-DA feedback session (this only occurred at the higher presentation rate). The second problem was convincing participants to come back for a third session, which was made worse by the end of the spring semester. Figure 12 shows the results for only the sham feedback sessions.

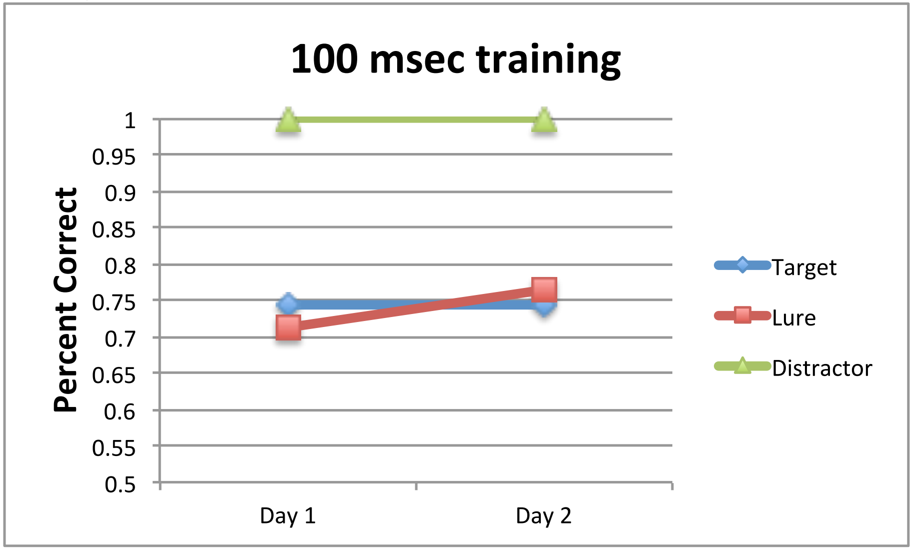

Figure 12

Summary

The overall goal of this research was to use a brain-computer interface (BCI) during a perceptual training paradigm to determine whether this could decrease the speed and increase the accuracy of making perceptual decisions compared to baseline conditions with no feedback. However, success was dependent on adapting several tasks and methods in a way that had not been tried before, and the difficulty in adapting these tasks and methods negatively impacted our research. In particular, the lack of turn-key system meant that we spent a large amount of time on software development. Although we had previous experience using BCI-lab to analyze data offline, when we started the project it was not capable of communicating in real-time with our Neuroscan hardware. In collaboration with members from other labs, we were able to write a Matlab plug-in for BCI lab that was compatible with the Neuroscan system (although bugs did become apparent during the higher presentation rates used during Spring 2015).

External Validity trade-off

Our initial goal was to use the one-shot paradigm, in which a single stimulus display is presented and the participant is required to respond as quickly and accurately as possible. Our initial review of the literature (Luo & Sajda, 2009; Johnson & Olshausen 2003) suggested that the effect size using the one-short technique was similar to the RSVP technique, and therefore it would be a good candidate for single-trial classification. Although we found grand-averaged ERPs to the targets using the one-shot paradigm, the signal-to-noise ratio was low enough that we were unable to get adequate single-trial classification. Therefore we turned to using the RSVP technique as used in a multitude of image-triage experiments. The downside is that the external validity is greatly decreased (the one-shot paradigm mimics what a person might see in a single glance, whereas the RSVP technique is completely artificial), and manual response times are no longer meaningful.

Initial RSVP results might have been overly task-dependent.

RSVP BCI might work best when the task is to detect the presence of a member of a high-level category. For example, Touryan and colleagues (2014) were able to consistently get stimulus-locked classification scores between .75 and .85 (A') when the task was to detect the presence of a high-level category (e.g. "stairs", "doors", "chairs"). Our initial high classification scores were for discriminating between the presence or absence of the target (e.g. T-72 target vs. empty distractor scene), and discrimination was slightly worse when trying to discriminate between trials that contained an armored target (e.g. "T-72") and an unarmored lure (e.g. school bus).

However, when we changed the task to discriminate between items that were part of the same high-level category (e.g. target is armored T-72 and lure is armored M1-Abrams, Training Experiment 4), then BCI classification fell to levels that were closer to 0.6 (A'), which is too inaccurate to use to provide meaningful feedback. This was particularly problematic, as using unarmored lures changes the task from "is there a T-72 present?" to "Is there an armored vehicle present?", and this was reflected in the high-behavioral accuracy. For an EEG-DA to be useful, it must provide classification that is more accurate than individuals can manage alone. In addition, high initial accuracy scores are problematic, as they allow for little room for training-based improvement.

Variability in individual subject BCI classification

In addition, there was a large effect of individual differences on BCI performance (see Figure 13). Our initial BCI performance after we switched to the RSVP technique yielded A' scores that were extremely encouraging (target vs. background A' ranged from 0.84-0.91, target vs. lure ranged from 0.73-0.83). However, we were not able to replicate this success with future participants. We spent a large part of our effort trying to track down causes (e.g. over-familiarity with stimuli) and markers (e.g. attentional blink magnitude) for these individual differences. Instead, it might have been more fruitful if we refined our presentation technique to try to maximize the p300 effect, perhaps by increasing the number of distractor scenes in each RSVP stream and decreasing the number of streams that contained a target. Both of these would have increased the rarity of the target, which is correlated with the magnitude of the p300 component, and might have allowed better BCI single-trial classification.

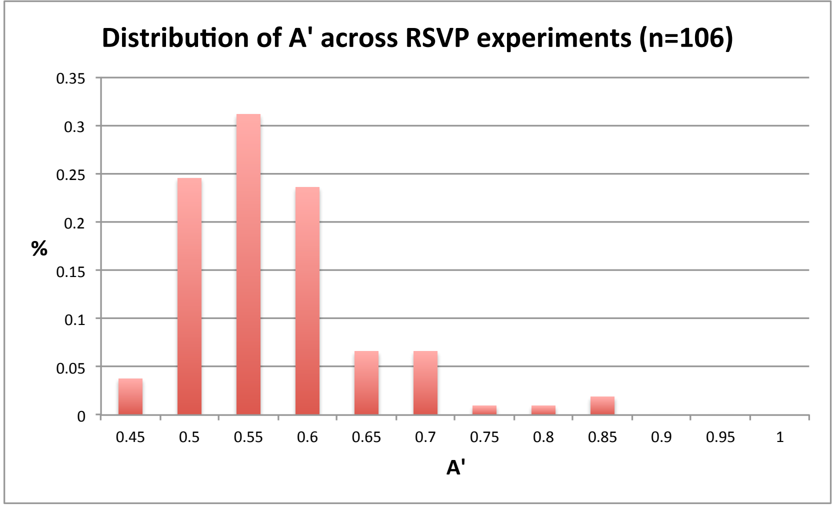

Figure 13

On the other hand, we did find individuals' classification solutions derived from prior session data worked as well on subsequent sessions data. These solutions did not transfer from person-to-person, but only worked well when they were used with subsequent sessions for the same individual. That is, a classification solution for person A's first session did not work for person B, but did work as well for person A's second session. This suggests that, had our presentation paradigm allowed us to get consistently high classification rates on the first session, that we would have been able to use solutions generated from previous sessions for training during subsequent sessions. This would have been key for using BCI as a training tool.

References

Chun, M. M., & Potter, M. C. (1995). A 2-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception & Performance*, *21*(1), 109-127.

Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision, 3(7)*, 499-512.

Luo, A., & Sajda, P. (2009). Comparing neural correlates of visual target detection in serial visual presentations having different temporal correlations. *Frontiers in human neuroscience*, *3*.

McArthur G, Budd T, Michie P. (1999) The attentional blink and P300. *Neuroreport* 10(17):3691-5.

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human Factors, 3,* 511-520.

Sajda, P., Gerson, A. D., Philiastides, M. G., & Parra, L. C. (2007). Single-trial analysis of EEG during rapid visual discrimination: Enabling cortically-coupled computer vision. *Towards brain-computer interfacing*, 423-44.

Sessa, P., Luria, R., Verleger, R., & Dell'Acqua, R. (2007). P3 latency shifts in the attentional blink: further evidence for second target processing postponement. *Brain research*, *1137*, 131-139.

Touryan, J., Marathe, A., & Ries, A. (2014). P300 variability during target detection in natural images: Implications for single-trial classification. *Journal of Vision*, *14*(10), 195-195.